

**Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
Пермский национальный исследовательский политехнический
университет**

Факультет	Электротехнический
Выпускающая кафедра:	Информационные технологии и автоматизированные системы
Направление подготовки:	09.04.01 «Информатика и вычислительная техника»
Профиль:	Автоматизированные системы обработки информации и управления
Квалификация:	Магистр
Дисциплина:	«Интеллектуальный анализ web-данных»

Отчёт по лабораторной работе №2

На тему: «Распространенные форматы слабоструктурированных данных»

Выполнил: студент группы АСУ4-22-1м

Попов К.М. (_____)

подпись

Проверил:

Ярулин Д. В., к.т.н. (_____)

подпись

Пермь, 2024

Задание

Требуется реализовать программу-транслятор между форматами HTML, (plain) XML, CSV, TSV и JSON.

На вход подается документ в любом из перечисленных форматов и указанный пользователем желаемый выходной формат.

На выходе — корректный документ в указанном формате.

Можно пользоваться встроенными парсерами форматов для загрузки и выгрузки.

Теория

HTML - «язык [гипертекстовой](#) разметки») — стандартизированный язык гипертекстовой разметки документов для просмотра [веб-страниц](#) в [браузере](#). Веб-браузеры получают HTML документ от сервера по протоколам [HTTP/HTTPS](#) или открывают с локального диска, далее интерпретируют код в интерфейс, который будет отображаться на экране монитора.

Пример:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<html>
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
  <title>Пример веб-страницы</title>
</head>
<body>
  <h1>Заголовок</h1>
  <!-- Комментарий -->
  <p>Первый абзац.</p>
  <p>Второй абзац.</p>
</body>
</html>
```

CSV (от [англ.](#) *Comma-Separated Values* — значения, разделённые запятыми) — текстовый формат, предназначенный для представления табличных данных. Строка таблицы соответствует строке текста, которая содержит одно или несколько полей, разделённых запятыми.

Пример:

```
QuotaAmount,StartDate,OwnerName,Username
150000,2016-01-01,Chris Riley,trailhead9.ub20k5i9t8ou@example.com
150000,2016-02-01,Chris Riley,trailhead9.ub20k5i9t8ou@example.com
150000,2016-03-01,Chris Riley,trailhead9.ub20k5i9t8ou@example.com
150000,2016-01-01,Harold Campbell,trailhead14.jibpbwvuy67t@example.com
150000,2016-02-01,Harold Campbell,trailhead14.jibpbwvuy67t@example.com
150000,2016-03-01,Harold Campbell,trailhead14.jibpbwvuy67t@example.com
150000,2016-01-01,Jessica Nichols,trailhead19.dlfxj2goytkp@example.com
150000,2016-02-01,Jessica Nichols,trailhead19.dlfxj2goytkp@example.com
150000,2016-03-01,Jessica Nichols,trailhead19.dlfxj2goytkp@example.com
150000,2016-01-01,Catherine Brown,trailhead16.kojyepokybge@example.com
150000,2016-02-01,Catherine Brown,trailhead16.kojyepokybge@example.com
150000,2016-03-01,Catherine Brown,trailhead16.kojyepokybge@example.com
150000,2016-01-01,Kelly Frazier,trailhead7.zdcsy4ax10mr@example.com
150000,2016-02-01,Kelly Frazier,trailhead7.zdcsy4ax10mr@example.com
150000,2016-03-01,Kelly Frazier,trailhead7.zdcsy4ax10mr@example.com
150000,2016-01-01,Dennis Howard,trailhead4.wfokpckfroxp@example.com
150000,2016-02-01,Dennis Howard,trailhead4.wfokpckfroxp@example.com
150000,2016-03-01,Dennis Howard,trailhead4.wfokpckfroxp@example.com
```

TSV ([англ. *tab separated values*](#) — значения, разделённые табуляцией) — [текстовый формат](#) для представления таблиц баз данных. Каждая запись в таблице — это строка текстового файла. Каждое поле записи отделяется от других с помощью символа [табуляции](#), точнее горизонтальной табуляции. TSV — это форма более общего формата [DSV](#) — значения, разделённые разделителем.

Пример:

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	I. setosa
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
4.6	3.1	1.5	0.2	I. setosa
5.0	3.6	1.4	0.2	I. setosa

JSON ([англ. *JavaScript Object Notation*](#)) — текстовый формат обмена данными, основанный на *JavaScript*. Но при этом формат независим от JS и может использоваться в любом языке программирования.

Пример:

```
{
  "glossary": {
    "title": "example glossary",
    "GlossDiv": {
      "title": "S",
      "GlossList": {
        "GlossEntry": {
          "ID": "SGML",
          "SortAs": "SGML",
          "GlossTerm": "Standard Generalized Markup Language",
          "Acronym": "SGML",
          "Abbrev": "ISO 8879:1986",
          "GlossDef": {
            "para": "A meta-markup language, used to create markup languages such as DocBook.",
            "GlossSeeAlso": ["GML", "XML"]
          },
          "GlossSee": "markup"
        }
      }
    }
  }
}
```

Практика

Перейдём к практике.

Использованная библиотека — pandas.

В качестве архитектура взаимодействия методов программы выбрана «звезда». С помощью библиотеки pandas мы можем преобразовать каждый формат в каждый формат. Всего можно сделать: $5^2 - 5 = 20$ преобразований

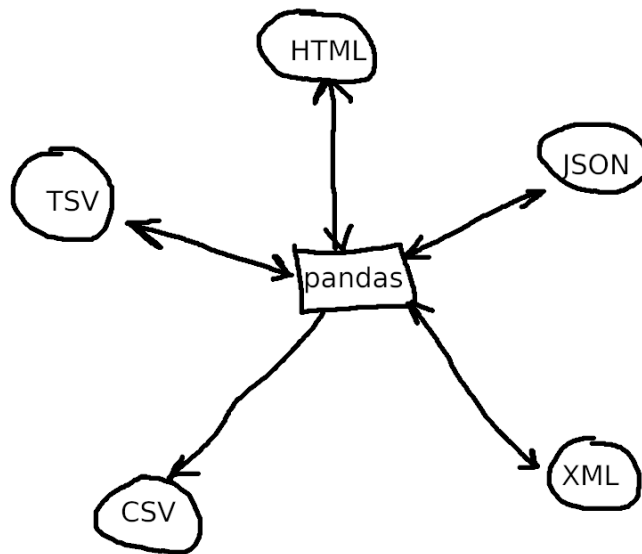
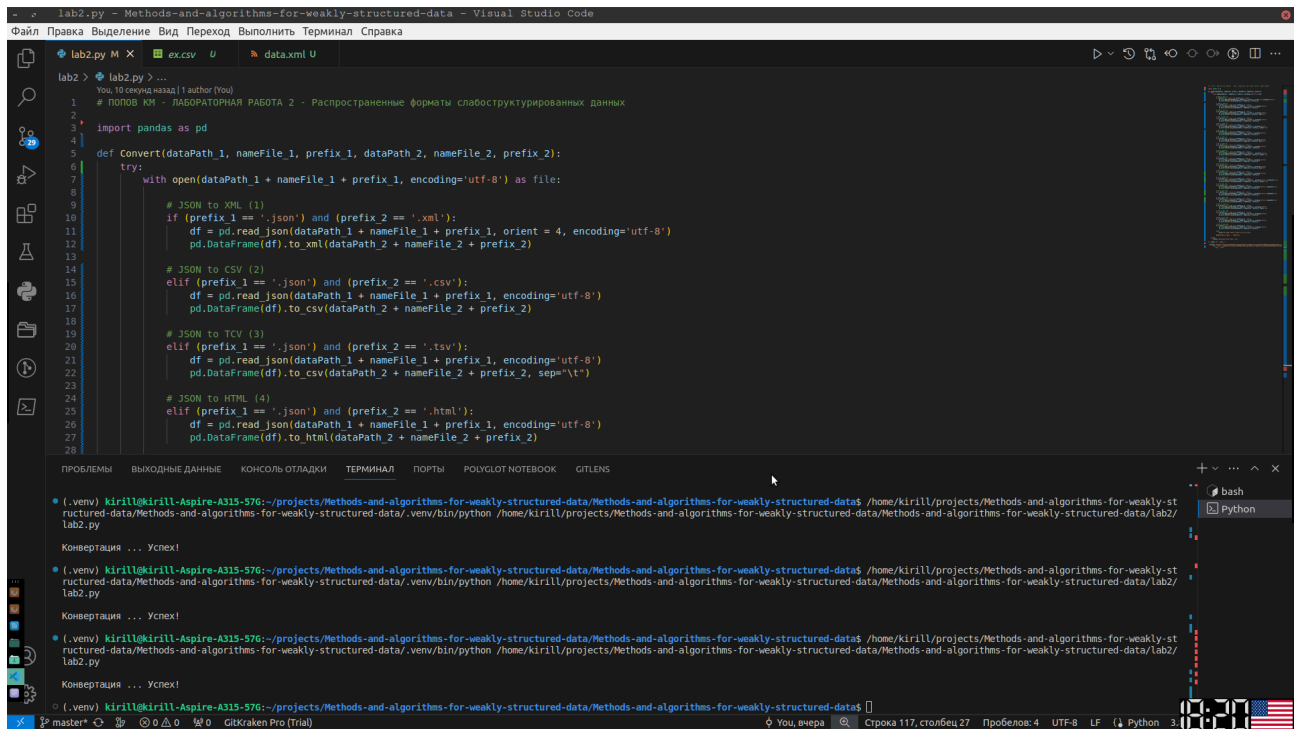


Рисунок 1 - Архитектура

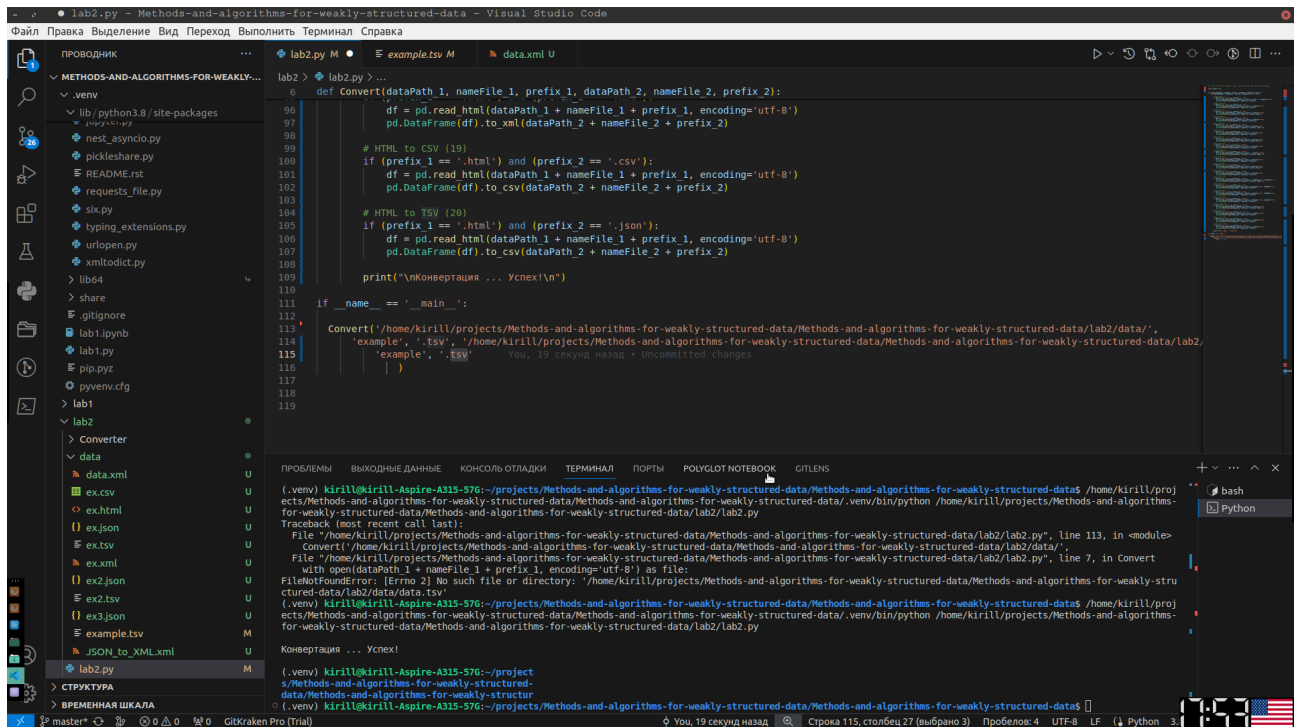
В исходном коде реализован 1 метод — `Convert(...)`. В его параметрах указывается путь до исходного файла с его расширением и путь до нового файла с выбранным расширением.

На рис. 2 и 3 приведены скриншоты работы программы. На них можно увидеть и результаты запуска.



The screenshot shows the Visual Studio Code editor with the file `lab2.py` open. The code defines a function `Convert` that takes `dataPath_1`, `nameFile_1`, `prefix_1`, `dataPath_2`, `nameFile_2`, and `prefix_2` as arguments. It uses `pandas` to read JSON data from `dataPath_1` and convert it to various formats (XML, CSV, TSV, HTML) based on the `prefix_2` parameter. The terminal output shows the successful execution of the script, with messages like "Конвертация ... Успех!" (Conversion ... Success!).

Рисунок 2 - Скриншот работы программы. Часть 1



This screenshot shows the same Visual Studio Code environment, but with a different part of the `lab2.py` file visible. The code now includes a section for converting HTML data to CSV or TSV. The terminal output shows an error message: `FileNotFoundError: [Errno 2] No such file or directory: '/home/kirill/projects/Methods-and-algorithms-for-weakly-structured-data/Methods-and-algorithms-for-weakly-structured-data/lab2/data/data.tsv'`. This indicates that the file specified in the code does not exist.

Рисунок 3 - Скриншот работы программы. Часть 2

На рис.4, 5, 6, 7, 8 приведены конвертированные данные.

```
thms-for-weakly-structured-data - Visual Studio Code
полнить Терминал Справка

lab2.py M ex.csv U data.xml U X

lab2 > data > data.xml
1  <?xml version='1.0' encoding='utf-8'?>
2  <data>
3    <row>
4      <index>0</index>
5      <shape>square</shape>
6      <degrees>360</degrees>
7      <sides>4.0</sides>
8    </row>
9    <row>
10     <index>1</index>
11     <shape>circle</shape>
12     <degrees>360</degrees>
13     <sides/>
14   </row>
15   <row>
16     <index>2</index>
17     <shape>triangle</shape>
18     <degrees>180</degrees>
19     <sides>3.0</sides>
20   </row>
21 </data>
```

Рисунок 4 - HTML

```
полнить Терминал Справка

lab2.py M ex.csv U X data.xml U

lab2 > data > ex.csv > data
1  ,level_0,index,shape,degrees,sides
2  0,0,0,square,360,4.0
3  1,1,1,circle,360,
4  2,2,2,triangle,180,3.0
5
```

Рисунок 5 - CSV

ПравкаВыделениеВидПереходВыполнитьТерминалСправка

lab2.py M ex.html U X data.xml U

lab2 > data > ex.html > table.dataframe > tbody > tr > td

```
1 <table border="1" class="dataframe">
2   <thead>
3     <tr style="text-align: right;">
4       <th></th>
5       <th>Unnamed: 0</th>
6       <th>level_0</th>
7       <th>index</th>
8       <th>shape</th>
9       <th>degrees</th>
10      <th>sides</th>
11    </tr>
12  </thead>
13  <tbody>
14    <tr>
15      <th>0</th>
16      <td>0</td>
17      <td>0</td>
18      <td>0</td>
19      <td>square</td>
20      <td>360</td>
21      <td>4.0</td>
22    </tr>
23    <tr>
24      <th>1</th>
25      <td>1</td>
26      <td>1</td>
27      <td>1</td>
28      <td>circle</td>
29      <td>360</td>
30      <td>NaN</td>
31    </tr>
32    <tr>
33      <th>2</th>
34      <td>2</td>
35      <td>2</td>
36      <td>2</td>
37      <td>triangle</td>
38      <td>180</td>
39      <td>3.0</td>
40    </tr>
41  </tbody>
42 </table>
```

Рисунок 6 - HTML


```
ыполнить Терминал Справка
lab2.py M ex.json U X data.xml
lab2 > data > {} ex.json > ...
1 {
2   "index":{
3     "0":0,
4     "1":1,
5     "2":2
6   },
7   "shape":{
8     "0":"square",
9     "1":"circle",
10    "2":"triangle"
11  },
12  "degrees":{
13    "0":360,
14    "1":360,
15    "2":180
16  },
17  "sides":{
18    "0":4.0,
19    "1":null,
20    "2":3.0
21  }
22 }
```

Рисунок 7 - JSON

```
ыполнить Терминал Справка
lab2.py M ex4.tsv U X data.xml U
lab2 > data > ex4.tsv > data
1 index shape degrees sides
2 0 0 square 360 4.0
3 1 1 circle 360
4 2 2 triangle 180 3.0
5
```

Рисунок 8 - TSV

Заключение

В результате выполнения лабораторной работы №2 была реализована программа преобразования форматов файлов XML, CSV, TSV и JSON.

Приведены скриншоты работы программы.

Приведены результаты конвертирования форматов.